TOPIC SEGMENTATION IN ASR TRANSCRIPTS USING BIDIRECTIONAL RNNS FOR CHANGE DETECTION

Imran Sehikh[*]

TCS Innovation Labs - Mumbai, India
imran.as@tcs.com

Dominique Fohr, Irina Illina

Université de Lorraine, LORIA, UMR 7503, France
Inria, Villers-lès-Nancy, France
CNRS, Vandoeuvre-lès-Nancy, France
{dominique.fohr, irina.illina}@loria.fr

ABSTRACT

Topic segmentation methods are mostly based on the idea of lexical cohesion, in which lexical distributions are analysed across the document and segment boundaries are marked in areas of low cohesion. We propose a novel approach for topic segmentation in speech recognition transcripts by measuring lexical cohesion using bidirectional *Recurrent Neural Networks* (RNN). The bidirectional RNNs capture context in the past and the following set of words. The past and following contexts are compared to perform topic change detection. In contrast to existing works based on sequence and discriminative models for topic segmentation, our approach does not use a segmented corpus nor (pseudo) topic labels for training. Our model is trained using news articles obtained from the internet. Evaluation on ASR transcripts of French TV broadcast news programs demonstrates the effectiveness of our proposed approach.

Index Terms— topic segmentation, recurrent neural networks

1. INTRODUCTION

The problem of topic segmentation, to automatically breakdown a text document into topically coherent segments, has been studied for a long time. With the increase in multimedia content on the internet, there has been an interest to extend topic segmentation to audio-video documents. Multimedia documents like broadcast news programs, meeting recordings, telephone conversations and lectures commonly consist of information on more than one topic. For example, broadcast news present events related to politics, economy, sports, weather and so on. Automatic segmentation of such documents, into coherent segments, is required by several down stream tasks such as topic detection and tracking [1], summarisation, named entity extraction and for multimedia indexing and organisation [2].

Approaches to topic segmentation are based on the idea of lexical cohesion [3]. Some of these methods analyse the lexical distribution across the document and mark segment boundaries in areas of low cohesion. This includes the original *TextTiling* algorithm [4] and its extensions using lexical chains [5], semantic/topic space representations [6, 7] and Laplacian Eigenmaps [8]. Another set of segmentation methods try to cluster together neighbouring areas instead of directly looking for topic boundaries. This includes the prominent *C99* algorithm [9] and its extensions using semantic/topic representations [10, 11, 12]. Alternative methods based on generative probabilistic models have also been proposed for topic segmentation. These include extensions of classical probabilistic topic models to incorporate topic changes and boundaries [13, 14, 15, 16], and approaches which directly model words in each topic segment as draws from a corresponding multinomial language model [17, 18, 19].

Most topic segmentation approaches were originally tried on textual resources or manual transcriptions of spoken resources. Unlike on text documents, topic segmentation on transcriptions obtained from an *Automatic Speech Recognition* (ASR) system cannot readily exploit sentence boundaries. In this regard, some works have used a fixed block of words as an unit [8, 20] while others have relied on pauses in speech [21]. Interestingly, prosodic cues and automatic sentence segmentation techniques have been leveraged in other works [22, 23, 24]. Performance of topic segmentation on ASR transcripts is also affected by word errors from the ASR. To reduce the effect of word errors, the use of ASR confidence measures and lattices have been proposed [25, 26]. Discriminative features from the speech signal, speaker patterns and news structure have shown to improve performance on spoken documents [3, 5, 21].

In this paper, we focus on lexical cohesion based topic segmentation on the ASR 1-best hypothesis. In contrast to the previous works, we propose a novel approach based on *Recurrent Neural Networks* (RNN). Recently RNNs have been shown to effectively model sequences like text, speech, music

[*] This work was performed while the author was a member of the Multispeech team of Université de Lorraine, Inria and CNRS.

and videos, and have given state-of-the-art results in several sequence and temporal classification tasks [27]. Their ability to model long term context and the potential to train them discriminatively motivates us to try RNNs for the task of topic segmentation. More specifically, we aim to capture lexical cohesion using a bi-directional RNN with *Long Short-Term Memory* (LSTM) cells [28, 27]. The bi-directional model would capture context from the past and the following set of words. These past and following contexts can then be compared to perform topic change detection.

Previous works on topic segmentation have adopted sequence modelling approaches based on *Hidden Markov Models* (HMM) [13, 17, 22, 29], and also discriminatively trained models including *Conditional Random Fields* (CRF) [30], deep feed forward neural network with HMM [29] and *Support Vector Machines* with sliding windows [26]. However, RNNs are better than fixed size windows and HMMs at exploiting contextual information [27]. Moreover, these works required a segmented training corpus and used (pseudo) topic labels of these segments for training their models (although pseudo labels were obtained in an unsupervised manner). As opposed to this, our proposed model is trained using news articles obtained from the internet.

The main contributions of this paper are (a) bi-directional RNNs for topic segmentation in ASR transcripts, which eliminate the need for an initial segmentation based on heuristics like fixed size windows or silence, (b) methods for discriminative training of bi-directional RNNs for topic segmentation without the need of a segmented training corpus and/or (pseudo) topic labels. We consider the application of segmentation of ASR transcripts of French broadcast news. The proposed model is evaluated on concatenated broadcast news videos as well as on real television news programs, and is compared to classical methods. The rest of the paper is organised as follows. Section 2 presents the idea behind our model and its architecture. It further includes a discussion on how our model is trained. Section 3 presents our experiment setup including the corpora, model configurations and evaluation methodology. This is followed by the segmentation results and a discussion in Section 3.4 and the conclusion in Section 4.

2. TOPIC CHANGE DETECTION WITH BI-DIRECTIONAL RNN

Topic segmentation can also be seen as a problem of topic change detection. A typical approach to mark a change point in a sequence is to take a window (of points) on either sides of a supposed change point and compare the adjacent windows using suitable features and representations. The comparison of adjacent windows gives a measure of whether the two windows belong to same the class or not. Following such a computation for all the points in the sequence, a similarity graph is obtained in which most of the peaks/valleys correspond to actual change points in the sequence. TextTiling [4] based algorithms for topic segmentation closely resemble this methodology.

Our proposed topic segmentation model is also based on a similar idea. The input is a sequence of words hypothesised by an ASR system and topic segmentation is to be performed on this 1-best ASR hypothesis. Words in the ASR hypothesis are represented using word embeddings. Instead of using fixed windows of words, RNNs are used to model the long term topic context and to perform change detection. More specifically we choose a bi-directional RNN [31] with LSTM cells. The bi-directional RNN allows to model context from the past and the following set of words, and it could be trained to measure topic cohesion between these contexts.

2.1. Model architecture

Our model consists of a layer of bi-directional RNN with LSTM cells. To understand the functioning of our model, a bi-directional LSTM-RNN is depicted in Figure 1.

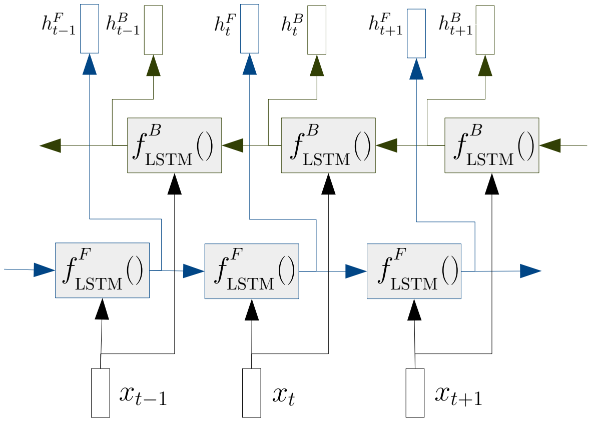


Fig. 1. Bi-directional RNN with LSTM cells

The inputs to our model are word embeddings [32] corresponding to words in the ASR 1-best hypothesis. These embeddings are denoted as $(x_1, ...x_{t-1}, x_t, x_{t+1}..., x_N)$, where x_t represents the current word and N is the length of the ASR hypothesis. Each word embedding is input to the forward and backward LSTM-RNNs, represented in Figure 1 by a chain of blocks of $f^F_{\text{LSTM}}()$ and $f^B_{\text{LSTM}}()$. The hidden layer activations for the forward LSTM-RNN at time t are denoted as h^F_t and those for the backward LSTM-RNN are denoted as h^B_t. Thus,

$$h^F_t = f^F_{\text{LSTM}}(x_t, h^F_{t-1}) \quad (1)$$

$$h^B_t = f^B_{\text{LSTM}}(x_t, h^B_{t+1}) \quad (2)$$

where, $f^F_{\text{LSTM}}()$ and $f^B_{\text{LSTM}}()$ represent the operations inside an LSTM cell with forget gates (refer [28, 27] for details). Hidden layer activations of the forward and backward LSTM-RNNs are transformed using a fully connected feed forward layer as follows:

$$c_t^F = h_t^F.W_{seg} + b_{seg} \tag{3}$$
$$c_t^B = h_t^B.W_{seg} + b_{seg} \tag{4}$$

where W_{seg} and b_{seg} are weight and bias parameters of the feed forward layer. Then the outputs corresponding to forward and backward LSTM-RNN at each t are compared as:

$$s_t = g(c_t^F.c_t^B) \tag{5}$$

where . denotes a dot product. This dot product compares the similarity between the topic context until t, as captured in c_t^F, and the topic context following t, which is captured in c_t^B. Thus $\{s_t\}_{t=1:N}$ represents the topic context similarity across the ASR hypothesis and this similarity should be minimum at the segment boundaries. The similarity calculation also involves an output function $g()$. We present two different possibilities for the output function $g()$, in Section 2.3.

2.2. Training without topic labels

Approaches based on sequence models [13, 17, 22] and discriminatively trained models [26, 29] require a topic segmented corpus, as well as (pseudo) topic labels for these segments, during training. These previous works have tackled this issue by automatically generating topic labels for segments of text in the training corpus. To obtain topic labels, unsupervised clustering is performed on the text segments and each resulting cluster is marked with a different topic. A similar technique can be employed for training our model with bi-directional RNNs. However, in this work we present an alternative approach which does not rely on a segmented corpus and topic labels.

Our training set consists of news articles crawled from a news website, as detailed in Section 3.1. Each article contains text around a particular news event. To train our model, two or more news articles are randomly chosen and concatenated. Then the training objective is to mark the boundary between the concatenated articles as a topic change point. The assumption is that the training set has a significantly large number of news articles and from a wide time period (ranging over months). With such a training set, it is very less likely for a generated training sample to contain two or more news segments which are on the same news event and adjacent to each other. Even otherwise an unsupervised topic model, like *Latent Dirichlet Allocation* (LDA) [33], can be built on the training set and topic distributions of adjacent segments can be compared to avoid concatenation of articles on same news events. However, a detailed analysis on selection of training set samples is not in the scope of this paper and we form our training set using news articles spread over a period of 6 months.

2.3. Output layer and training criteria

As discussed in Section 2.2, the training objective is to mark the boundary between the concatenated training set samples. Recall that in the last layer of our model, in (5), the similarity value should be minimum at the boundary. Accordingly, we present two possibilities for the choice of the output function $g()$, which can achieve this training objective.

2.3.1. Softmin Function

$$g(d_{t*}) = \frac{e^{-d_{t*}}}{\sum_{t=1}^{N} e^{-d_t}} \tag{6}$$

The softmin function will give a high output probability to a low dot product similarity value ($d_t = c_t^F.c_t^B$). For training with softmin, a cross-entropy cost function can be used. The cost is calculated as:

$$\mathfrak{L} = -log(s_{t'}) \tag{7}$$

where t' is the position of the topic change in input text. The softmin function followed by the cross-entropy cost will bring a discriminative capability to the segmentation model by maximising the likelihood of the true segmentation point ($t*$, known at the time of training), as compared to all the points ($t = 1, 2, ...N$). During test the high output values will correspond to the hypothesised topic change points.

2.3.2. Flipped sigmoid function

$$g(d_t) = 1 - \text{sigmoid}(d_t) = \frac{1}{1 + e^{d_t}} \tag{8}$$

This function assigns a value close to zero to a low dot product similarity value ($d_t = c_t^F.c_t^B$) and a value close to 1 to a high dot product similarity. During training a binary cross-entropy cost function will be used, which is given as:

$$\mathfrak{L} = -\sum_{\substack{t=1:N \\ t \neq t'}} log(s_t) - \sum_{t'} log(1 - s_{t'}) \tag{9}$$

where t' are positions of topic change in the input text. During test the output value at each t is regraded as the probability of change point at the corresponding word in the input text. This output function, unlike the softmin function, is (a) independent of the length (N) of the input text and (b) allows training using input samples with more than one change point (for e.g. by concatenating more than two articles during training).

It should be noted that a softmax function and a standard sigmoid function can be used in place of the softmin and flipped sigmoid functions. However, our choice is motivated by the fact that dot product values in (5) represent similarity values and should be minimum at the topic change point. Results from our initial experiments, using different output functions, also supported this argument.

3. EXPERIMENTS SETUP

3.1. Experiment corpora

We use two set of corpora in our experiments. The first set consists of 24,000 news articles from French newspaper *L'Express*, and about 3000 news articles and 3000 news videos from the French website of the *Euronews* TV channel. These news articles and videos appeared during the period January - June 2014. More details about these datasets can be found in [34]. Each article/video in this dataset contains news on a particular event.

For training our model we concatenate randomly chosen articles from the *L'Express* dataset. Our validation set consists of 3000 samples formed by concatenating 2 articles at a time from *Euronews* text articles. Our first test set, referred as the *Euronews Test Set*, consists of 3000 samples with each sample obtained by concatenating 2 to 7 videos from *Euronews*. (News videos from *Euronews* are 2 to 5 min in duration.) For training and test, the punctuation marks are removed and all words are converted to lower case, as it would be with ASR transcripts.

Apart from the *Euronews Test Set*, we evaluate our models on a real test set consisting of 20 news programs that appeared on the French TV channel *TV5* during February 2017. This test set is referred as the *TV5 Test Set*. On average a program is 12 minutes long and consists of 3 to 10 segments.

3.2. Model configurations and training

The model with the softmin function in the output layer, as discussed in Section 2.3.1, will be denoted as 'RNN-SMIN'. It is trained using samples obtained by concatenating 2 articles at a time from the *L'Express* corpus. Model with the flipped sigmoid function in the output layer, as discussed in Section 2.3.2, is trained with samples obtained by concatenating 2 to 4 articles at a time from the *L'Express* corpus. This model configuration will be denoted as 'RNN-FSIG'.

As inputs to the RNN-SMIN and RNN-FSIG, model we use pre-trained 200 dimensional Skip-gram word embeddings[1], which were trained on '.fr' domain websites. These word embeddings were not updated during training so that the model generalises to unseen words in the test set. In addition, we applied 50% dropout at the output of the word embedding layer to achieve generalisation and avoid overfitting. The word dropout also adds robustness to ASR errors, as we have shown earlier in [35]. For model training, we used mini-batch stochastic gradient descent with ADADELTA [36]. An early stopping criteria was used, which stops the model training when the error on the validation set starts increasing.

3.3. Evaluation setup

The topic segmentation models will be evaluated on the reference and ASR transcriptions of the *Euronews Test Set*, and on the ASR transcriptions of the *TV5 Test Set*. The ASR transcriptions are obtained from our French ASR system based on DNN-HMM acoustic models. Our ASR gives a word error rate of 16.4% on the *Euronews Test Set*.

We compare the performance of our models with two baseline topic segmentation methods. Our first baseline is the TopicTiling algorithm proposed in [37]. TopicTiling is based on the classical TextTiling algorithm for topic segmentation [4] and uses topic assignments from a LDA topic model. Our second baseline is the classical C99 algorithm for topic segmentation [9], improved with representations learnt using *Latent Semantic Analysis* (LSA) [10, 11]. This baseline will be denoted as 'C99-LSA'.

For the C99-LSA and TopicTitling baseline methods, 200 dimensional LSA and LDA models were trained on the same training corpus that was used to train our RNN models. Stopwords were removed while training these baseline models, as well as the proposed models. Additionally lemmatisation was applied in case of baseline models. C99-LSA operates on sentence level representations. Since ASR transcripts do not have punctuations we used fixed blocks of words instead. Similarly, TopicTitling was used with fixed blocks of words as basic units. The block size was tuned using the validation set.

For comparison of the proposed and baseline methods, we use the standard topic segmentation evaluation measures P_k and WD [3]. They indicate the probability of segmentation error, with a lower value indicating a better performance. As compared to P_k, WD penalises false alarms. We also calculate the Precision (P) and Recall (R) [3] on the segmentation results, and report the $F1$ score which was calculated as $F1 = 2(P \times R)/(P + R)$. A window of 25 words on each side of a true segmentation point was used to label a hypothesised segmentation point as true positive or false positive.

3.4. Topic segmentation results and discussion

Table 1 presents the topic segmentation results on the Euronews Test Set. Results on both reference and ASR transcriptions are shown. Note that the training corpus, used for the baseline as well as the proposed models, and the Euronews Test Set consist of news from the same time period (see Section 3.1). So it is likely that they contain similar topics and hence it is a matched train-test condition. Results in Table 1 show that the C99-LSA method gives the lowest P_k and WD errors. However, our proposed model RNN-FSIG gives the best $F1$ score and its P_k and WD is quite close to that of C99-LSA. For the Euronews Test Set our RNN-SMIN model performs only better than TopicTitling. Further analysis of the results on this test set revealed that both our proposed models have high precision (about 0.8) as compared to that of C99-

[1] obtained from `http://fauconnier.github.io`

LSA (at about 0.65). RNN-SMIN model showed a relatively lower recall.

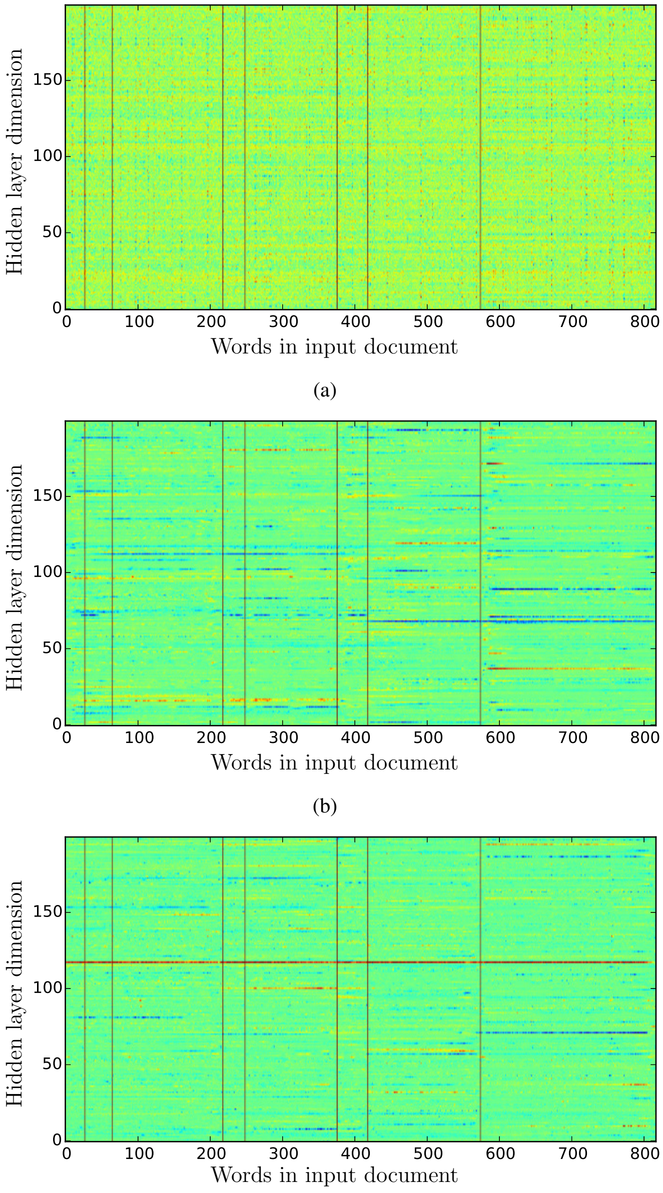
Topic segmentation results on the TV5 Test Set are presented in Table 2. Only ASR transcription results are shown (due to lack of reliable reference transcriptions). Note that this test set is from a different time period than our training set (see Section 3.1) and hence there is a mis-matched train-test conditions in terms of contents and topics. Results in Table 2 show that our proposed models perform better than the baseline models, which seem to be affected by the mis-matched train-test conditions. On contrary to results on the Euronews Test Set, RNN-SMIN gives the best results and RNN-FSIG has a performance similar to that of RNN-SMIN.

Table 1. Topic segmentation error (P_k, WD) and $F1$ score on the Euronews Test Set (with matched train-test conditions). Best results are highlighted in bold.

	Reference			ASR		
	P_k	WD	$F1$	P_k	WD	$F1$
TopicTitling	0.44	0.49	0.55	0.45	0.49	0.54
C99-LSA	**0.21**	**0.25**	0.75	**0.22**	**0.26**	0.73
RNN-SMIN	0.24	0.32	0.69	0.25	0.32	0.66
RNN-FSIG	0.21	0.26	**0.79**	0.24	0.28	**0.78**

Table 2. Topic segmentation error (P_k, WD) and $F1$ score on ASR transcripts of the TV5 Test Set (with mis-matched train-test conditions). Best results are highlighted in bold.

	P_k	WD	$F1$
TopicTitling	0.38	0.45	0.53
C99-LSA	0.29	0.35	0.60
RNN-SMIN	**0.26**	**0.34**	**0.66**
RNN-FSIG	0.26	0.34	0.64



(a)

(b)

(c)

Fig. 2. Visualisations of hidden layer activations in the RNN-FSIG model for the ASR transcript of a news program from TV5 French TV channel. (a) input word embeddings (b) LSTM output activations from the forward RNN (c) LSTM output activations from the backward RNN.

To further demonstrate the effectiveness of our proposed approach and to highlight the learning in our models we present Figure 2. It shows visualisations of hidden layer activations in the RNN-FSIG model, for an ASR transcript in the TV5 Test Set. Visualisation of the 200 dimensional input word embeddings are shown in Figure 2(a) and visualisations of the LSTM output activations from the forward and backward LSTM-RNN are shown in Figure 2(b) and Figure 2(c), respectively. In each of the figures, the X-axes represent the words in the ASR transcript and the Y-axes represent the 200 dimensions. The vertical lines in each visualisation represent the true topic boundaries, corresponding to the 8 segments, in the program. We can see that the word embeddings do not show any clear patterns for the topic segments but the activations from the forward and backward LSTM-RNN show patterns (appearing as coloured horizontal lines) on certain dimensions. For example, activations in Figure 2(b) clearly higlight segments 2, 4, 5, 7 and 8 (from left to right). Similarly, projections in Figure 2(c) show activations for segments 5, 6, 7 and 8. These activations and their patterns demonstrate that the LSTM-RNNs have learned about topic coherence and acquired the ability to perform topic segmentation.

4. CONCLUSIONS

We proposed a novel approach for topic segmentation which measures lexical and topic cohesion using bidirectional Recurrent Neural Networks. The bi-directional RNNs captured context in the past and the following set of words, and performed topic change detection by comparing the past and following contexts. These models were trained discriminatively by concatenating news articles from the internet. Evaluation on ASR transcripts of French TV news programs showed that our RNN models can perform better than the C99-LSA and TopicTiling baseline methods. Our models achieved best $F1$ scores throughout, due to higher precision rates, and performed well in mis-matched conditions where the baselines started lacking. With the use of standard word embeddings our model can be readily adapted to topic segmentation in different domains.

5. ACKNOWLEDGEMENTS

This work is funded by the ContNomina project supported by the French National Research Agency (ANR) under contract ANR-12-BS02-0009. Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria, CNRS, RENATER and other Universities and organisations (https://www.grid5000.fr).

6. REFERENCES

[1] James Allan, *Topic Detection and Tracking: Event-Based Information Organization*, Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[2] Lin shan Lee and B. Chen, “Spoken document understanding and organization,” *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 42–60, Sept 2005.

[3] Matthew Purver, *Spoken Language Understanding*, chapter Topic Segmentation, pp. 291–317, John Wiley & Sons, Ltd, 2011.

[4] Marti A. Hearst, “Texttiling: Segmenting text into multi-paragraph subtopic passages,” *Comput. Linguist.*, vol. 23, no. 1, pp. 33–64, Mar. 1997.

[5] Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing, “Discourse segmentation of multi-party conversation,” in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, 2003, ACL ’03, pp. 562–569.

[6] Andrew Olney and Zhiqiang Cai, “An orthonormal basis for topic segmentation in tutorial dialogue,” in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005, pp. 971–978.

[7] Martin Riedl and Chris Biemann, “Text segmentation with topic models,” *Journal for Language Technology and Computational Linguistics*, vol. 27, no. 1, pp. 47–69, 2012.

[8] L. Xie, L. Zheng, Z. Liu, and Y. Zhang, “Laplacian eigenmaps for automatic story segmentation of broadcast news,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 276–289, Jan 2012.

[9] Freddy Y. Y. Choi, “Advances in domain independent linear text segmentation,” in *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, 2000, pp. 26–33.

[10] Freddy Y. Y. Choi, Peter Wiemer-Hastings, and Johanna Moore, “Latent semantic analysis for text segmentation,” in *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, Lillian Lee and Donna Harman, Eds., 2001, pp. 109–117.

[11] Yves Bestgen, “Improving text segmentation using latent semantic analysis: A reanalysis of choi, wiemer-hastings, and moore (2001),” *Comput. Linguist.*, vol. 32, no. 1, pp. 5–12, Mar. 2006.

[12] Riedl, M., Biemann, and C., “Text segmentation with topic models,” *Journal for Language Technology and Computational Linguistics (JLCL)*, vol. 27, no. 1, pp. 47–70, Aug. 2012.

[13] David M. Blei and Pedro J. Moreno, “Topic segmentation with an aspect hidden markov model,” in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2001, pp. 343–348.

[14] Matthew Purver, Thomas L. Griffiths, Konrad P. Körding, and Joshua B. Tenenbaum, “Unsupervised topic modelling for multi-party spoken discourse,” in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, 2006, pp. 17–24.

[15] Lan Du, Wray Buntine, and Mark Johnson, “Topic segmentation with a structured topic model,” in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, June 2013, pp. 190–200.

[16] Lan Du, John K Pate, and Mark Johnson, "Topic segmentation with an ordering-based topic model," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015, pp. 2232–2238.

[17] J. P. Yamron, I. Carp, L. Gillick, S. Lowe, and P. van Mulbregt, "A hidden markov model approach to text segmentation and event tracking," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 1998, vol. 1, pp. 333–336 vol.1.

[18] Masao Utiyama and Hitoshi Isahara, "A statistical model for domain-independent text segmentation," in *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, 2001, pp. 499–506.

[19] Jacob Eisenstein and Regina Barzilay, "Bayesian unsupervised topic segmentation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2008, pp. 334–343.

[20] Lei Xie, Jia Zeng, and Wei Feng, "Multi-scale texttiling for automatic story segmentation in chinese broadcast news," in *Proceedings of the 4th Asia Information Retrieval Conference on Information Retrieval Technology*, 2008, pp. 345–355.

[21] Abdessalam Bouchekif, Graldine Damnati, Yannick Estve, Delphine Charlet, and Nathalie Camelin, "Diachronic semantic cohesion for topic segmentation of tv broadcast news.," in *INTERSPEECH*, 2015, pp. 2932–2936.

[22] Gökhan Tür, Dilek Hakkani-Tür, Andreas Stolcke, and Elizabeth Shriberg, "Integrating prosodic and lexical cues for automatic topic segmentation," *Computational Linguistics*, vol. 27, no. 1, pp. 31–57, 2001.

[23] Gina-Anne Levow, "Prosody-based topic segmentation for mandarin broadcast news," in *Proceedings of HLT-NAACL 2004: Short Papers*, Stroudsburg, PA, USA, 2004, HLT-NAACL-Short '04, pp. 137–140, Association for Computational Linguistics.

[24] Andrew Rosenberg, Mehrbod Sharifi, and Julia Hirschberg, "Varying input segmentation for story boundary detection in english, arabic and mandarin broadcast news," in *INTERSPEECH*, 2007, pp. 2589–2592.

[25] Camille Guinaudeau, Guillaume Gravier, and Pascale Sébillot, "Enhancing lexical cohesion measure with confidence measures, semantic relations and language model interpolation for multimedia spoken content topic segmentation," *Computer Speech & Language*, vol. 26, no. 2, pp. 90 – 104, 2012.

[26] Mehryar Mohri, Pedro Moreno, and Eugene Weinstein, "Discriminative topic segmentation of text and speech," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 533–540.

[27] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A Critical Review of Recurrent Neural Networks for Sequence Learning," *ArXiv e-prints*, May 2015.

[28] Felix A. Gers, Jürgen A. Schmidhuber, and Fred A. Cummins, "Learning to forget: Continual prediction with lstm," *Neural Comput.*, vol. 12, no. 10, pp. 2451–2471, Oct. 2000.

[29] Jia Yu, Xiong Xiao, Lei Xie, Eng Siong Chng, and Haizhou Li, "A dnn-hmm approach to story segmentation," in *INTERSPEECH*, 2016, pp. 1527–1531.

[30] Xiaoxuan Wang, Lei Xie, Mimi Lu, Bin Ma, Engsiong Chng, and Haizhou Li, "Broadcast news story segmentation using conditional random fields and multimodal features," *IEICE Transactions*, vol. 95-D, no. 5, pp. 1206–1215, 2012.

[31] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, Nov 1997.

[32] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*, pp. 3111–3119. Curran Associates, Inc., 2013.

[33] David M Blei, Andrew Y Ng, and Michael I Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[34] Imran Sheikh, Irina Illina, and Dominique Fohr, "How diachronic text corpora affect context based retrieval of oov proper names for audio news," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, may 2016, pp. 3851–3855.

[35] I. Sheikh, D. Fohr, I. Illina, and G. Linars, "Modelling semantic context of oov words in large vocabulary continuous speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 598–610, March 2017.

[36] M. D. Zeiler, "ADADELTA: An Adaptive Learning Rate Method," *ArXiv e-prints*, Dec. 2012.

[37] Martin Riedl and Chris Biemann, "Topictiling: A text segmentation algorithm based on lda," in *Proceedings of ACL 2012 Student Research Workshop*, 2012, ACL '12, pp. 37–42.